

Critical fluctuations in proteins native states

Qian-Yuan Tang¹, Yang-Yang Zhang¹, Jun Wang¹, Wei Wang¹, and Dante R Chialvo²

¹ National Laboratory of Solid State Microstructure, Department of Physics,
and Collaborative Innovation Center of Advanced Microstructures,
Nanjing University, Nanjing 210093, China and

² Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET), Godoy Cruz 2290, Buenos Aires, Argentina
(Dated: January 15, 2016)

We study a large data set of protein structure ensembles of very diverse sizes determined by nuclear magnetic resonance. By examining the distance-dependent correlations in the displacement of residues pairs and conducting finite size scaling analysis it was found that the correlations and susceptibility behave as in systems near a critical point implying that, at the native state, the motion of each amino acid residue is felt by every other residue up to the size of the protein molecule. Furthermore certain protein's shapes corresponding to maximum susceptibility were found to be more probable than others. Overall the results suggest that the protein's native state is critical, implying that despite being posed near the minimum of the energy landscape, they still preserve their dynamic flexibility.

Protein molecules are formed by large unbranched chains of amino acids, which turn into a complex folded shape as free energy is minimized. It is this highly specific three-dimensional folded structure, known as native state, that makes the protein capable of performing its biological function [1]. Proteins carry out their functions by switching from one shape to another, even transiently, as for instance when it recognizes and binds with another molecule. To achieve such performance the structure of the native state must be very susceptible to sense the signal and switch to another shape, but also be stable enough to warrant reproducibility. It is well known that these apparently contradictory demands are exhibited by systems near a critical point because of the coexistence of maximum susceptibility and long range correlations [2–6].

These views are discussed on a number of recent reports emphasizing different aspects of critical fluctuations in the protein equilibrium dynamics. This includes the geometric properties [7], the slowness in relaxation in the dynamics of large biomolecules [8], the role of their low-frequency global modes [9, 10] in the proteins' functional dynamics, the overlap of the large-scale conformational change in allosteric transitions and the low frequency normal modes[11], the role of the water surrounding the molecule [12], as well as the near-critical states emerging in the sequential correlations of protein families [3].

Although it is often recognized that the available data seems still far from being the ideal to test for criticality, we propose here an approach to investigate this issue. We use a large number of protein structure ensembles determined by solution nuclear magnetic resonance (NMR). Since each ensemble contains different structures of the same protein, the basic idea is to assume that each of structures can be seen as a hypothetical instantiation of the spontaneous conformational changes that the protein exhibit through time. By examining the

distance-dependent correlations in the displacement of residue pairs and conducting finite size scaling analysis it is shown that the correlations and susceptibility behave as expected in systems near a critical point. The results imply that at the native state, the motion of each and every amino acid residue is felt by every other residue, up to the size of the protein molecule.

Fluctuations and correlations: Data and definitions. The dataset analyzed contains 7678 protein structure ensembles with not less than 10 different structures from the protein data bank (PDB)[13] (see complete list and details in the Supp. Info.). For each structural ensemble, we selected one configuration as a reference state, then by doing 3D structure alignment, the degrees of freedoms related to the translational and rotational motion are removed. As Fig.1(A) shows one conformation (colored in red) is set as the reference state, and the other conformations were aligned to that reference state. After the alignment, the displacement of every atom from the reference state was computed (Fig.1(B)). The calculations are based on the python package “ProDy” [14].

For simplification, we mainly focus on the C_α traces of the proteins. For a protein molecule made up of N amino acid residues, for all residue pairs i and j ($1 \leq i, j \leq N$), the distance r_{ij} between the two residues is approximated as the distance between the two C_α 's, and the average correlation C_{ij} (the elements of the covariance matrix) of the displacement of the two residues i and j is approximated by the average inner product of the displacement of two C_α atoms in the two residues.

In this manner it is constructed the covariance matrix (Fig.1(C)), $C_{ij} = \langle \Delta \vec{r}_i \cdot \Delta \vec{r}_j \rangle$, where $\Delta \vec{r}_i$ is the displacement from the average configuration of the C_α atom in residue i and $\Delta \vec{r}_j$ is the similar displacement for atom j . From the covariance C_{ij} , the orientational correlation of residue pairs $\phi_{ij} = C_{ij} / \sqrt{C_{ii} \cdot C_{jj}}$ is obtained. Here, C_{ii} and C_{jj} are the auto-correlation of the displacement of residue i and j , which are proportional to the B factors

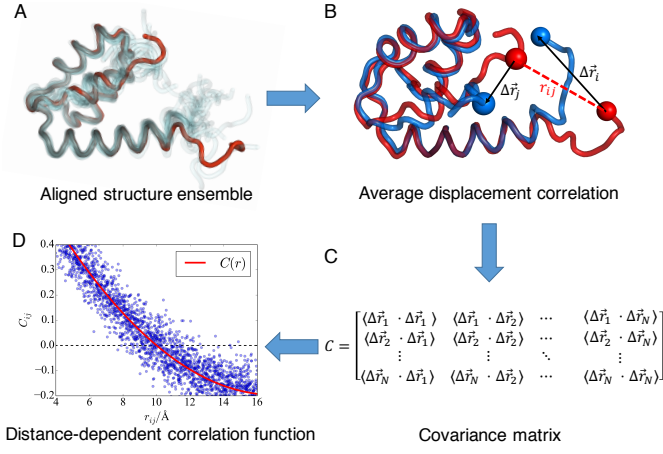


FIG. 1. (A) Aligned C_α traces of the structure ensemble of a protein molecule (PDB code: 2KQ6) determined by NMR. (B) An illustration of the definition of average displacement correlation $C_{ij} = \langle \Delta \vec{r}_i \cdot \Delta \vec{r}_j \rangle$. (C) The elements in a covariance matrix C . (D) The scattering plot of the r_{ij} and C_{ij} and the averaged distance-dependent correlation function $C(r)$.

(root mean square fluctuations) of the C_α atoms.

For our protein data, C_{ij} (and ϕ_{ij}) is roughly a function $C(r)$ of the distance r_{ij} (as shown in Fig.1(D)). Thus, for any given protein molecule, one calculate the average covariance and crosscorrelation for all the residue pairs that have similar distance between two C_α atoms ($r_{ij} \approx r$), allowing to define the distance-dependent covariance $C(r)$ and crosscorrelation $\phi(r)$ of single protein molecules as:

$$C(r) = \frac{\sum_{i,j} C_{ij} \delta(r - r_{ij})}{\sum_{i,j} \delta(r - r_{ij})}; \phi(r) = \frac{\sum_{i,j} \phi_{ij} \delta(r - r_{ij})}{\sum_{i,j} \delta(r - r_{ij})}. \quad (1)$$

Moreover, to reveal the general properties in the dynamics of proteins with same sizes, the average distance-dependent correlation functions $C(r)$ and $\phi(r)$ could also be calculated by similarly averaging the C_{ij} and ϕ_{ij} pairs for the proteins with similar radius of gyration R_g .

Correlation length is proportional to protein size: As shown in Fig.2, for proteins with different R_g , the average crosscorrelation functions $\phi(r)$ exhibit a maximum (vertical dashed line in Fig.2(A) at $r = 3.8\text{\AA}$) which corresponds to the covalent bonds. Then, for distances between residues close to R_g , the correlation function crosses zero, which defines the correlation length ξ_ϕ of $\phi(r)$ (Fig.2(A)). (Since we focus on the crosscorrelation function, we just denote ξ_ϕ as ξ). Beyond such a distance, the average correlation function is not vanishing, but first decreases to a negative minimum and then eventually approaches zero again. We note that this behavior of the correlation function is quite robust across different proteins; it only differs at the very long lengths due to boundary effects where the specific shape of each pro-

tein dominates the behavior of the correlation function. This behavior means that within a protein (independently of its size), there is either strong correlation (short distance) or strong anticorrelation (large distance), but there is no region in which the correlation is consistently negligible. For all the proteins studied the correlation length ξ_ϕ and R_g are found to be approximately proportional, (see Fig.2(B)) which indicates that the residues of a protein are correlated to every other, independently of how large the protein is. It means that if we could perturb a single residue, the consequence of such perturbation could be sensed by the entire protein molecule.

Consequently, all the correlation functions computed at various sizes, can be rescaled by its gyration radius R_g (or alternatively by its correlation length ξ). As shown in Fig.2(C), all the curves collapse together after rescaling the distance between residues as r/ξ . Moreover, the distance-dependent covariance function $C(r)$ and the fluctuation of B-factors also can be successfully rescaled, which indicates that not only the orientational crosscorrelation but also the amplitude of the fluctuations exhibits such kind of scale-free behavior (as occurs with the velocity correlations in the case of birds' flocks described in Ref.[15]).

A more careful analysis reveals that the covariance correlations contain additional information about residue-residue interactions. For proteins with different sizes, different kinds of amino acid pairs would have different distance-dependent covariances, for example, as shown in Fig.2(D), for leucine-leucine pairs, the covariance would in average be smaller than that of other listed types of residue pairs. This is because, usually, hydrophobic residues are buried in the core of proteins so that the fluctuations would be small; and the zero points of the covariance are also slightly influenced by the type of amino acids, since glycine is a very small amino acid, the correlation length would be slightly deviated from the average. To refine the force field for coarse-grained models, one should take into consideration all the detailed residue specific interaction information, which is reflected in the distance-dependent covariance $C(r)$. However, for all types of residues pairs, the distance-dependent crosscorrelation $\phi(r)$ (as shown in Fig.2(D) inset) still keeps the scale-free correlation with a similar correlation length.

Additional hints from finite size scaling: The type of correlations described above resembles those observed in the collective behavior of a variety of biological systems [2, 3, 15, 16], in which correlations are amplified by the vicinity to some critical point in the parameters space. However, most often, the system size is very small respect to the thermodynamic limit, such that the value of the control parameter at which correlation and susceptibility peak depends on size. Thus in order to stay critical some inverse relation need to be found between

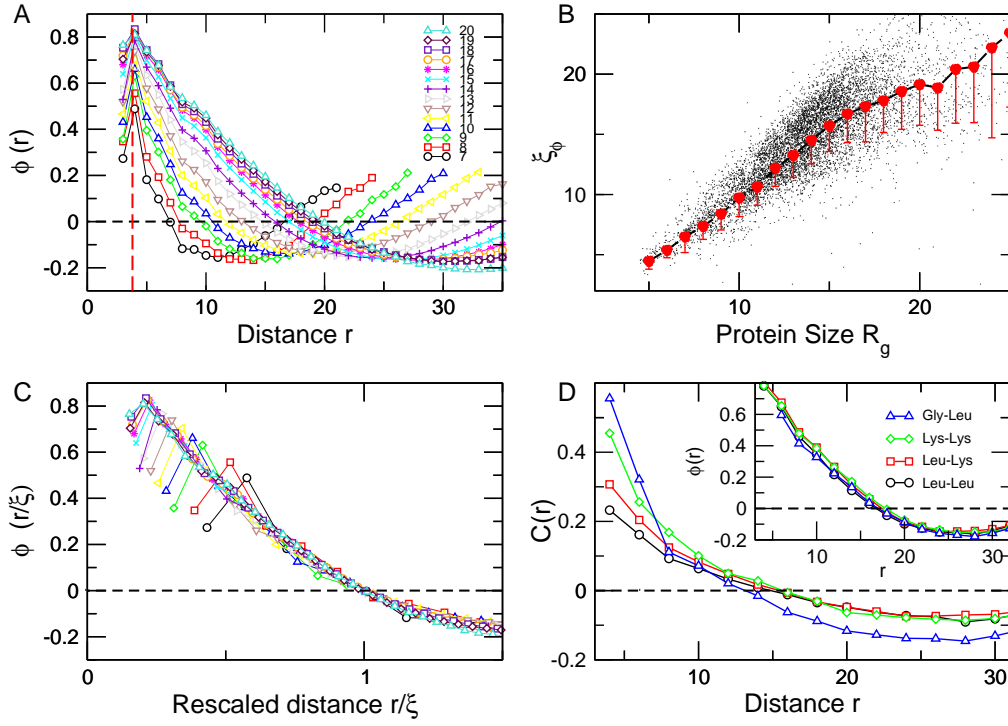


FIG. 2. (A) The distance-dependent crosscorrelation function $\phi(r)$ averaged for proteins with similar R_g . (B) Scattering plot of correlation length ξ_ϕ as a function of the average R_g , where the red symbols show the average $\xi = f(R_g)$ with the error bars denoting the standard deviation. (C) The scaling plot of $\phi(r/\xi)$. (D) The average distance-dependent covariance $C(r)$ and crosscorrelation function $\phi(r)$ (inset), for specific kinds of residues in proteins of $R_g \approx 15\text{\AA}$.

a (pseudo) control parameter and the system size. In turn, as the elegant demonstration of Attanasi et al.[16] shows, this finite-size scaling effect can be used to probe near-criticality.

Thus, we test such possibility for proteins of different sizes, performing similar finite-size scaling as was done in Ref. [16]. To start, for each protein, the susceptibility is defined as the summed crosscorrelation between residue pairs, that is:

$$\chi = \frac{1}{N} \sum_{i \neq j}^N \phi_{ij} \cdot \theta(\xi_\phi - r_{ij}). \quad (2)$$

Subsequently a dimensionless shape factor s is defined as the pseudo control parameter of the protein, $s = Na^3/(L_a L_b L_c)$, where $a = 3.8\text{\AA}$ is the size of a residue and L_a , L_b and L_c are the lengths of the principle axis of the protein ($L_a \leq L_b \leq L_c$). Such a parameter can also be understood as “packing density” because $L_a L_b L_c$ is proportional to the volume of an ellipsoid. For sphere-like protein molecules, the value of s is relatively large (densely packed, and solid-like), while for elongated chains (loosely packed, and polymer-like), $L_c = Na$, and $L_a = L_b = a$, thus $s = 1$.

As shown in Fig.3(A), the computation of the susceptibility χ for proteins of similar R_g reveals that the $\chi - s$

plot exhibits a series of maximum χ_m . Notice that when R_g increases the shape factor s for χ_m decreases, i.e., larger “critical” proteins seem to be more non-spherical than small ones. Also notice (in the inset) that susceptibility scales with protein size $\chi_m \sim R_g^{\gamma/\nu}$.

If the results correspond to (near) critical behavior then the following relations are expected to hold: $s \sim N^{-1/3\nu}$, and $\xi \sim R_g$, as well as $\chi \sim N^{\gamma/3\nu}$. Despite relatively large fluctuations the data exhibit scaling behavior as shown in the fittings of Fig.3(B-D). For $s \sim R_g^{-1/\nu} \sim N^{-\alpha/\nu}$, and $R_g \sim N^\alpha$ we get $\alpha = 0.34$ (Panel B), thus $1/\nu \approx 0.96 \approx 1.09$, $\nu \approx 1.04 \approx 0.9$. Panel E shows that for the relation $\chi \sim s^{-\gamma}$, $\gamma = 3.2$. Since $\chi \sim R_g^{\gamma/\nu} \sim N^{\alpha\gamma/\nu}$, leads in both cases $\nu \approx 1.1$ (Panel D). Taking integers we could consider that $\alpha = 1/3$, $\nu = 1$, $\gamma = 3$.

Interestingly, it seems as if nature favors certain “critical” proteins, such that large and small proteins end up “adjusting” their shape in the folding process such that they remain susceptible. The inset of Fig.3(E) shows that the most frequent shape factor corresponds approximately to the maximum susceptibility values. A more detailed analysis shows (see the Supp. Info) that for proteins of similar R_g value the most frequent s corresponds to the maximum χ (i.e., at the critical line) at that R_g .

The scale free nature of the correlations makes the

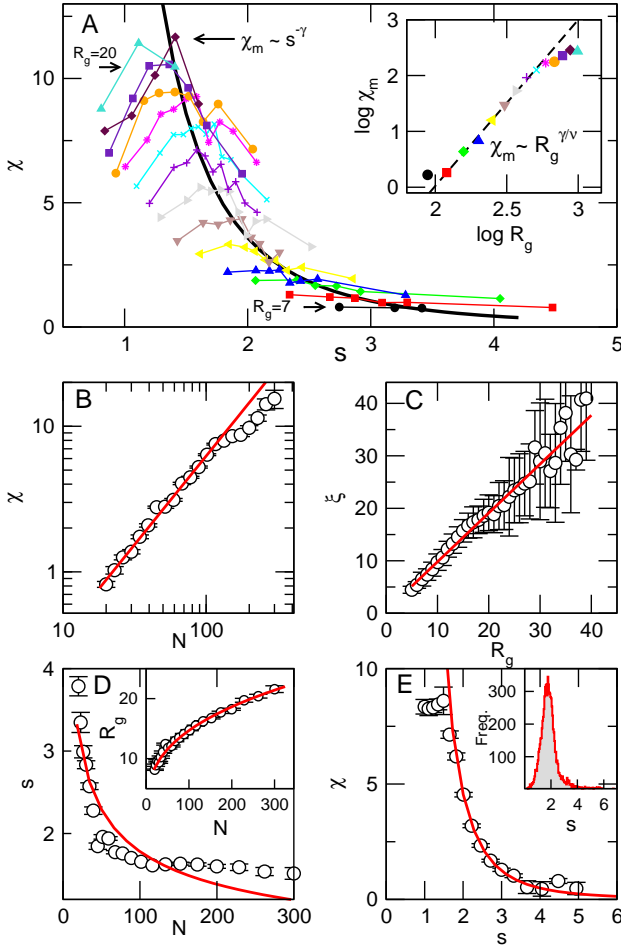


FIG. 3. Finite size scaling. (A) Susceptibility χ for proteins with different R_g as a function of the control parameter, s . Each curve is calculated for proteins of similar R_g . The peak of each proteins' susceptibility χ_m scales both with its shape (dark thick line) and size (inset). Same symbols and colors as in Fig. 2 indicating protein size, R_g . (B) Susceptibility, χ , as a function of number of residues, N . (C) Correlation length, ξ , as a function of protein size, R_g . (D) Control parameter s as a function of N . Inset shows R_g as a function of N . (E) Susceptibility, χ , as a function of s in main plot, and the distribution of s for all proteins in the inset.

distant residues well correlated with each other so that the global modes of the proteins could be easily excited. That is to say that even though the shape of the proteins varies widely from one to another, it seems that the folding process towards the native state builds a shape for the proteins which ensures the emergence of long-range correlations, which is needed to produce conformational changes as well as to keep memory of the current configuration.

Summarizing, the results show that the correlation length of the native state fluctuations is proportional to the gyration radius of the molecule, implying that the motion of any amino acid could influence all the others, up to the entire protein molecule. These results suggest that the proteins native states are not only posed near the minimum of the energy landscape, but also once there, they preserve the dynamic flexibility. In addition it is found that certain shapes are more probable, such that for any given protein size the folding process favors the shape with the maximum susceptibility (i.e., critical).

This work was supported by Natural Science Foundation of China (Grants 11334004, 11174133, 81421091) and National Basic Research Program of China (Grant No. 2013CB834100) and by CONICET of Argentina. Q-YT & DRC acknowledge the hospitality of the Max Planck Institute for the Physics of Complex Systems at Dresden (Germany). Email addresses: QYT: tangqianyan@gmail.com; JW: wangj@nju.edu.cn; WW: wangwei@nju.edu.cn; DRC: dchialvo@conicet.gov.ar.

- [1] J.R. Banavar and A. Maritan, *Annu. Rev. Biophys. Biomol. Struct.* **36** 261–80 (2007).
- [2] P. Bak, *How nature works: the science of self-organized criticality* (Springer, 2013).
- [3] T. Mora and W. Bialek, *J. of Stat. Phys.* **144**, 268 (2011).
- [4] A.R. Honerkamp-Smith, S.L. Veatch, S.L. Keller, *Biochimica et Biophysica Acta* **1788**, 53–63 (2009).
- [5] D.R. Chialvo, *Nature Physics* **6**, 744 (2010).
- [6] H. Chaté and M. Muñoz, *Physics* **7**, 120 (2014).
- [7] M.A. Moret, *Physica A* **390**, 3055–59 (2011).
- [8] H.P. Lu, L. Xun, X. S. Xie, *Science* **282**, 1877 (1998).
- [9] I. Bahar, A.R. Atilgan, M.C. Demirel, B. Erman, *Phys. Rev. Lett.* **80**, 2733 (1998).
- [10] I. Bahar, T.R. Lezon, L.-W. Yang, E. Eyal, *Ann. Rev. of Biophysics* **39**, 23 (2010).
- [11] L. Yang, G. Song, R.L. Jernigan, *Biophys. J.* **93**, 920 (2007).
- [12] A.J. Patel, P. Varilly, S.N. Jamadagni, M.F. Hagan, D. Chandler, S. Garde, *J. Phys. Chem. B* **116**, 2498–2503 (2012).
- [13] F.C. Bernstein, T.F. Koetzle, G.J. Williams, E.E. Meyer Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, *J. of Mol. Biol.* **112**, 535 (1977).
- [14] A. Bakan, L.M. Meireles, I. Bahar, *Bioinformatics* **27**, 1575 (2011).
- [15] A. Cavagna, A. Cimorelli, I. Giardina, G. Parisi, R. Santagati, F. Stefanini, and M. Viale, *PNAS* **107**, 11865 (2010).
- [16] A. Attanasi, A. Cavagna, L. Del Castello, I. Giardina, S. Melillo, L. Parisi, O. Pohl, B. Rossaro, E. Shen, E. Silvestri, et al., *Phys. Rev. Lett.* **113**, 238102 (2014).